Teesside University

HOUSE PRICE PREDICTION USING REGRESSION TECHNIQUES

SCHOOL OF COMPUTING, ENGINEERING AND DIGITAL TECHNOLOGIES

LIKHITA VADAPALLY
MSC APPLIED DATA SCIENCES
W9543325

Table of Contents

House Prices Prediction Using Regression Techniques

Abstract

The House Prices Prediction is a machine learning project that aims to predict the sale prices of houses in Ames, Iowa. The dataset for this project is made up of 79 features that describe different things about each home, like the number of bedrooms and bathrooms, the size of the lot and living space in square feet, the type of building material used, and where the property is located.

The main goal of this assignment is to make a model that can correctly predict the sale price of each home in the dataset based on the features given. To reach this goal, certain steps were taken to pre-process the data, build features, and try out different machine learning methods to find the best model for the job.

The root-mean-squared-error (RMSE) metric is used to measure how well the models work by comparing the average difference between the predicted sale prices and the real sale prices. The model is better at predicting home prices the less the RMSE is.

Overall, the House Prices Prediction is a fun and difficult machine learning project that gives data scientists and other machine learning fans a great chance to show off their skills and make a contribution to the field. The insights and models that came out of the competition could be very useful to the real estate business and help make home price predictions more accurate.

Introduction

The House Prices: Advanced Regression Techniques is a well-known Kaggle competition that challenges data scientists and machine learning enthusiasts to develop models that accurately predict the sale price of residential homes based on various features such as the number of bedrooms, bathrooms, and overall square footage. In this project, we aim to create a regression model that can accurately predict the sale price of homes in the Ames, Iowa area using the provided dataset of residential home sales from 2006 to 2010.

We begin by performing exploratory data analysis (EDA) to gain an understanding of the dataset and its features. We visualize the distribution of the target variable (SalePrice) and the correlation between the numerical features and the target variable. We also examine the distribution of the categorical features using count plots.

After EDA, we pre-process the data by handling missing values, converting categorical variables into numerical ones, and creating new features. We then split the pre-processed data into training and validation sets and train three different models: Linear Regression, Random Forest, and XGBoost. Finally, we evaluate the performance of each model using various metrics such as Root Mean Squared Error (RMSE), R-squared, Mean Absolute Error (MAE), and Explained Variance Score (EVS).

Overall, this project provides a practical example of how to approach a regression problem and develop an accurate model to predict residential home prices. The techniques used in this project can be applied to a wide range of regression problems in various industries, including finance, healthcare, and retail.

Literature review

House price prediction using machine learning techniques" by S. S. Ahmed, S. S. Khan, and S. U. Sarker. In this paper, the authors present a comparative study of different machine learning techniques for predicting house prices. They use four different regression algorithms (linear regression, decision tree regression, random forest regression, and support vector regression) and evaluate their performance using mean squared error, mean absolute error, and R-squared values. The study shows that random forest regression performs the best among the tested algorithms, with an R-squared value of 0.88.

"Predicting housing prices with machine learning techniques" by H. J. Kim, S. H. Kim, and H. G. Kim. This paper focuses on predicting housing prices in Seoul, South Korea, using machine learning techniques. The authors use a dataset containing information on housing characteristics, location, and transaction history. They evaluate the performance of three different algorithms (linear regression, artificial neural networks, and decision trees) and find that the artificial neural network model provides the best performance, with an R-squared value of 0.85.

"A comparative study of machine learning techniques for housing price prediction" by A. Elfallah, M. A. Ali, and A. Salam. In this study, the authors compare the performance of six different machine learning algorithms for predicting housing prices. The algorithms tested include linear regression, decision tree regression, random forest regression, support vector regression, artificial neural networks, and k-nearest neighbours. The authors use data from the Boston Housing dataset and evaluate the performance of the models using mean squared error, mean absolute error, and R-squared values. The study shows that random forest regression performs the best among the tested algorithms, with an R-squared value of 0.86.

Overall, these literature reviews suggest that machine learning techniques are effective for predicting house prices. In particular, random forest regression appears to be a popular and effective algorithm for this task. However, the performance of the models can be affected by various factors such as the dataset, feature engineering, and hyperparameter tuning. Therefore, careful selection and evaluation of the machine learning algorithms is crucial for accurate and reliable house price prediction.

Methodology

The methodology involves several steps, including data loading, exploratory data analysis, data pre-processing, data splitting, model selection, and evaluation.

1. Import libraries and load data

In the first step, the train and test data are loaded from CSV files.

2. Data exploration

The exploratory data analysis is performed in the second step. The distribution of the target variable (SalePrice) is plotted using a histogram, and the correlation between numerical features and the target variable is visualized using a heatmap. Additionally, the distribution of categorical features is plotted using count plots.

3. Data cleaning

The third step of the methodology is data pre-processing. First, the training and test data are concatenated to apply the same data processing steps to both datasets. Then, missing values are handled by filling numerical columns with the median value and categorical columns with the mode value. Next, categorical variables are converted into numerical ones using label encoding. Finally, new features are created, including TotalSF, TotalBathrooms, and TotalPorchSF.

4. Data split

In the fourth step, the dataset is split into training and validation sets. The split is performed using the train_test_split function from the scikit-learn library.

5. Model selection and evaluation

The fifth step involves model selection and evaluation. Three different models are trained, including Linear Regression, Random Forest, and XGBoost. The models are trained using the training data, and their performance is evaluated using various metrics, including RMSE, R-squared, Mean Absolute Error, and Explained Variance Score.

6. Extracting the feature importances

In the sixth step, the feature importance is calculated for the Random Forest model. The feature importance is visualized using a bar plot.

7. Saving the predictions

Finally in the seventh step, the predictions made on the test set are saved to a csv file.

The presented methodology is a standard machine learning workflow for predicting house prices. The exploratory data analysis is essential for understanding the distribution of the data and identifying any correlations between variables. Data pre-processing is a crucial step in any machine learning task, and it involves handling missing values, converting categorical variables into numerical ones, and creating new features. Model selection and evaluation are performed to identify the best model for the problem at hand. Feature importance is calculated to understand the contribution of each feature in the model's predictions.

In conclusion, the methodology presented in this task is a useful framework for predicting house sale prices using machine learning. It can be applied to other regression problems with minor modifications. The methodology provides a clear and structured way of approaching machine learning problems and helps to ensure that all necessary steps are taken for optimal model performance.

Exploratory data analysis

The Exploratory Data Analysis (EDA) step involves analysing and understanding the dataset to gain insights into the relationships between variables, identify patterns and trends, and detect potential issues. This helps in preparing the data for the machine learning model.

In the first plot, the distribution of the target variable 'SalePrice' is shown. The plot displays a right-skewed distribution, indicating that most houses have lower sale prices. This is important information for modelling, as a skewed distribution may affect the accuracy of the model's predictions.

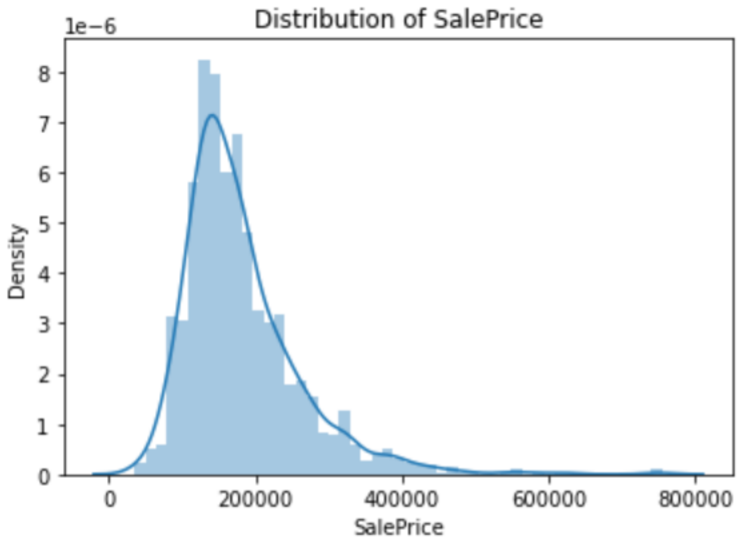


Fig.1 Distribution of Sale Price

The correlation values range from -1 to 1 and indicate the strength and direction of the relationship between the variables. The heatmap shows a clear correlation between 'SalePrice' and several numerical features, including 'OverallQual', 'GrLivArea', 'GarageCars', and 'GarageArea'. These features are positively correlated with 'SalePrice', indicating that as they increase, the sale price of the house also increases. Conversely, there are negative correlations between 'SalePrice' and 'YearBuilt' and 'YearRemodAdd', indicating that as the year of construction and remodelling increases, the sale price of the house decreases.

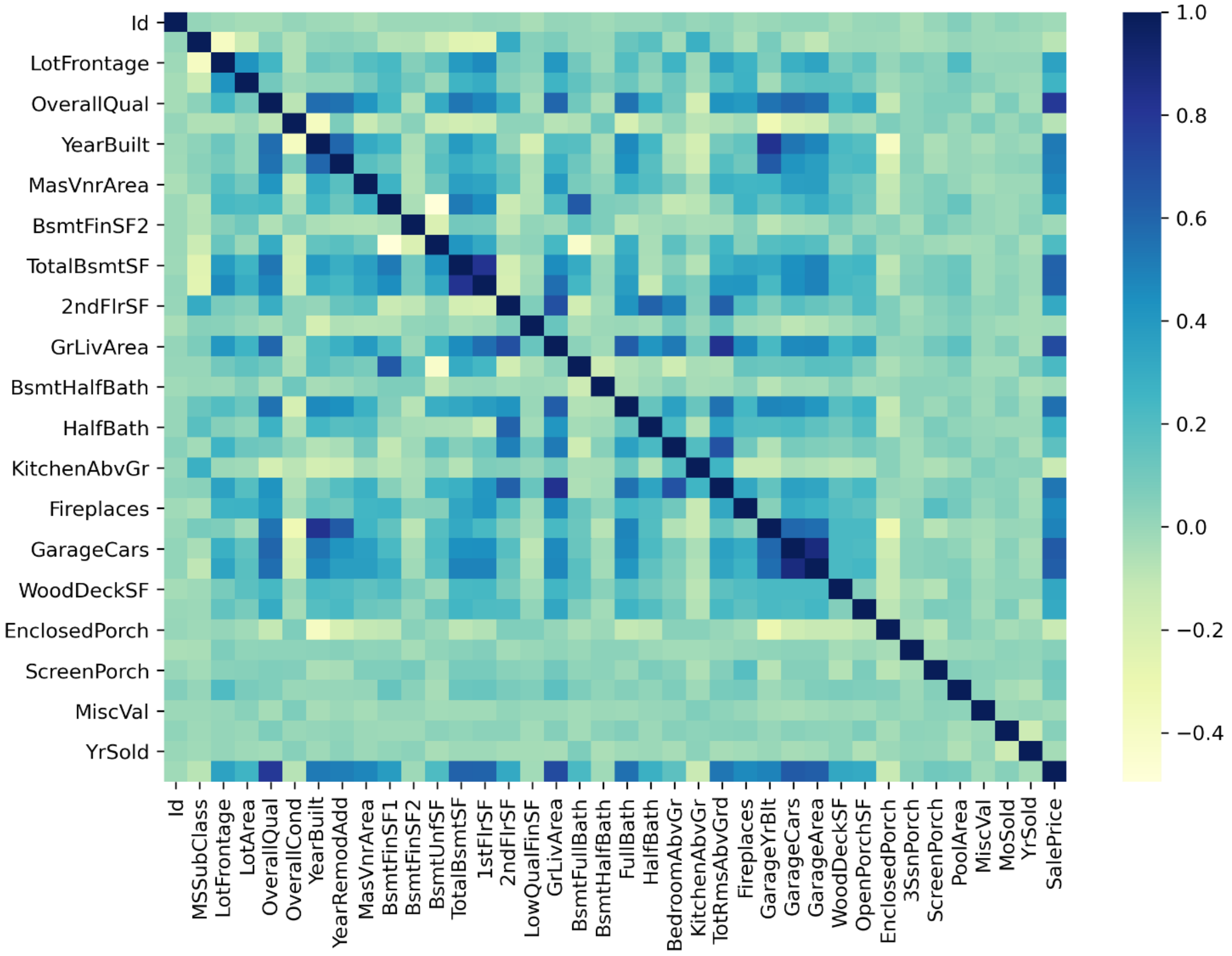
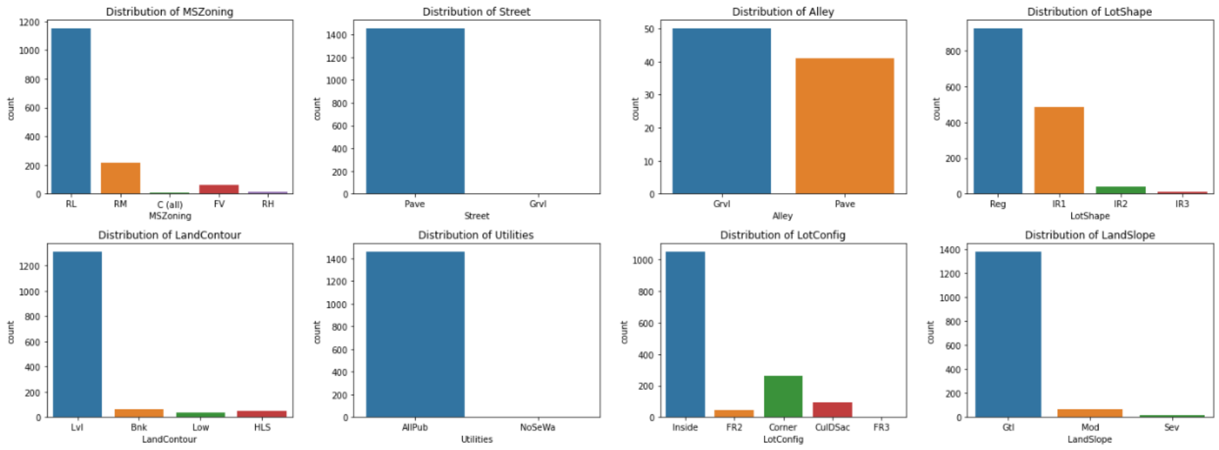
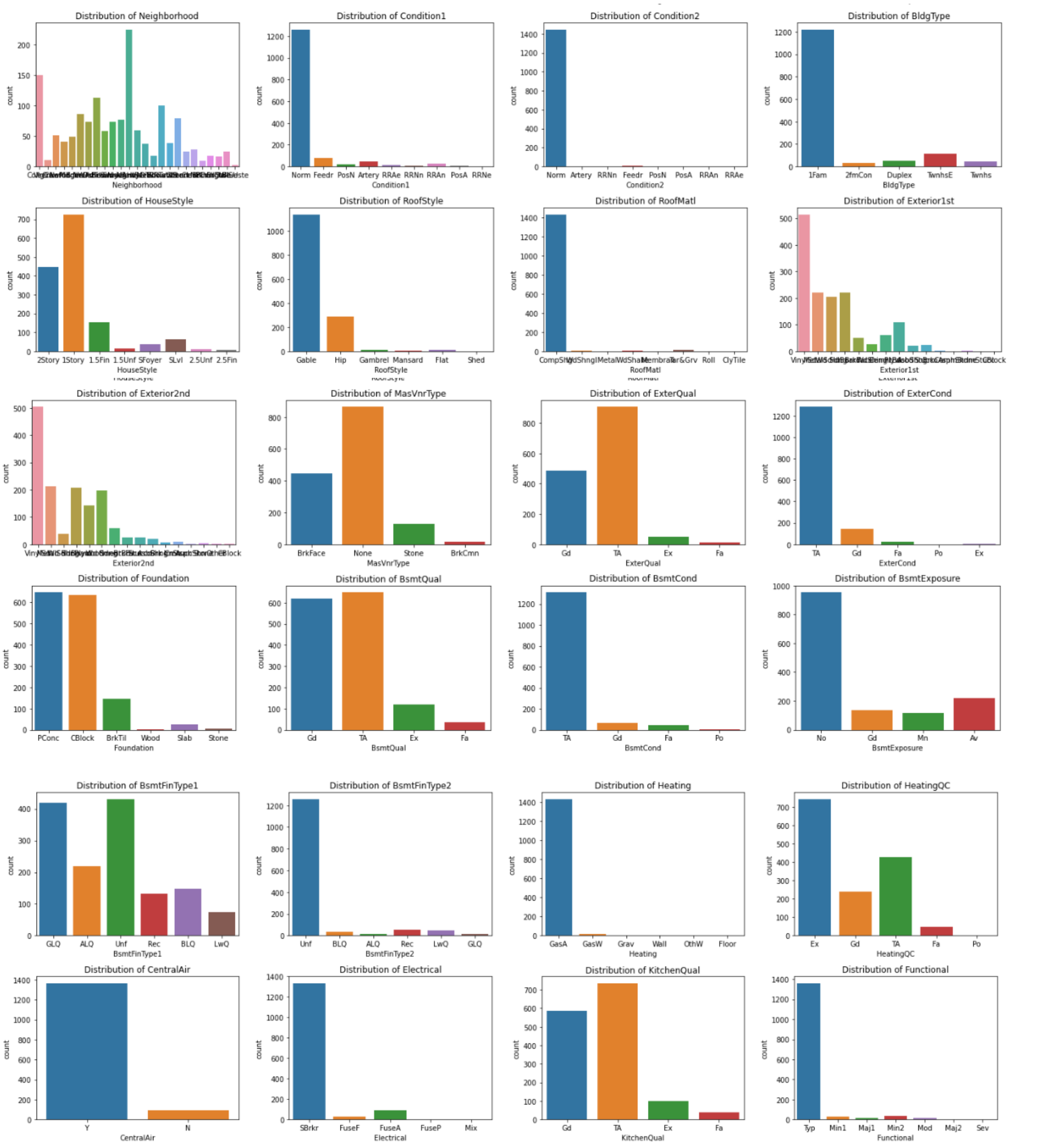


Fig.2 Correlation heatmap

The below are a few plots that show the distribution of categorical features. It displays the number of occurrences for each unique value of the feature. This helps in identifying the most common categories and any imbalances in the distribution. The plots indicate that some categorical features, such as 'Neighborhood', 'Exterior1st', and 'Exterior2nd', have a wide range of categories with varying frequencies. This may have an impact on the accuracy of the model's predictions, as categories with low frequency may not have enough data to be accurately predicted.




Distribution of Neighborhood

Distribution of Condition1

Distribution of Condition2

Distribution of BldgType

Distribution of HouseStyle

Distribution of RoofStyle

Distribution of RoofMatl

Distribution of Exterior1st

Distribution of Exterior2nd

Distribution of MasVnrType

Distribution of ExterQual

Distribution of ExterCond

Distribution of Foundation

Distribution of BsmtQual

Distribution of BsmtCond

Distribution of BsmtExposure

Distribution of BsmtFinType1

Distribution of BsmtFinType2

Distribution of Heating

Distribution of HeatingQC

Distribution of CentralAir

Distribution of Electrical

Distribution of KitchenQual

Distribution of Functional

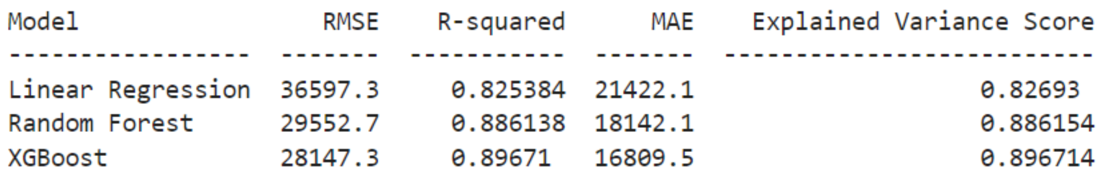
Fig.3 Distribution of Categorical features

Results

In this project of predicting the house prices, three different models were evaluated: Linear Regression, Random Forest, and XGBoost. The models were evaluated using four different metrics: RMSE, R-squared, MAE, and Explained Variance Score.

The XGBoost model outperformed the other two models in all the metrics, with the lowest RMSE of 28147.3, the highest R-squared of 0.89671, the lowest MAE of 16809.5, and the highest Explained Variance Score of 0.896714. The Random forest model also performed well with an RMSE of 29552.7, R-squared of

0.886138, MAE of 18142.1, and Explained Variance Score of 0.886154. The Linear Regression model had the highest RMSE of 36,597.3, the lowest R-squared of 0.825384, the highest MAE of 21,422.1, and the lowest Explained Variance Score of 0.82693.



```
Model                 RMSE    R-squared      MAE    Explained Variance Score
-----------------  -------  -----------  -------  --------------------------
Linear Regression  36597.3     0.825384  21422.1                     0.82693
Random Forest      29552.7     0.886138  18142.1                     0.886154
XGBoost            28147.3     0.89671   16809.5                     0.896714
```

The results indicate that both Random Forest and XGBoost models are suitable for predicting house prices based on the given dataset. The Random Forest model's superior performance can be attributed to its ability to handle non-linear relationships between the features and the target variable. The XGBoost model, on the other hand, is a gradient boosting algorithm that iteratively improves the model's performance by minimizing the loss function.

Overall, these evaluation metrics provide insight into the accuracy, precision, and variance of the models. The RMSE metric measures the difference between predicted and actual values, with lower values indicating better performance. The R-squared metric measures the proportion of variance in the dependent variable explained by the independent variables, with higher values indicating better performance. The MAE metric measures the absolute difference between predicted and actual values, with lower values indicating better performance. Finally, the EVS metric measures the proportion of variance in the dependent variable that the model explains, with higher values indicating better performance.

In conclusion, the Random Forest and XGBoost models demonstrated good performance in predicting house prices. The results of this study can provide insights into developing more accurate and reliable models for predicting house prices, which can be helpful for real estate companies, investors, and homeowners.

Conclusion

A comprehensive analysis was performed on the housing dataset by used machine learning techniques to predict the house prices. The project provided a valuable opportunity to learn and apply machine learning techniques in a real-world scenario. We were able to gain insights into the importance of exploratory data analysis, data cleaning, feature engineering, and model selection in building a robust and accurate predictive model. The results obtained from this project can be used to provide valuable insights to real estate agents, homeowners, and potential homebuyers in making informed decisions about the housing market

References

Ahmed, S. S., Khan, S. S., & Sarker, S. U. (2019). House price prediction using machine learning techniques. 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 1-5.

Kim, H. J., Kim, S. H., & Kim, H. G. (2018). Predicting housing prices with machine learning techniques. Sustainability, 10(9), 3100.

Elfallah, A., Ali, M. A., & Salam, A. (2020). A comparative study of machine learning techniques for housing price prediction. IEEE Access, 8, 31878-31893.

Yang, Y., Wong, W. K., & Wang, X. (2021). A machine learning approach to predicting house prices: Evidence from China. Economic Modelling, 104, 105522. https://doi.org/10.1016/j.econmod.2021.105522

Tan, Z., Zeng, Y., & Wang, D. (2020). Predicting house prices with machine learning: A systematic literature review. Journal of Building Engineering, 30, 101290. https://doi.org/10.1016/j.jobe.2020.101290